

Thermodynamic reversibility in feedback processes

JORDAN M. HOROWITZ and JUAN M. R. PARRONDO

Departamento de Física Atómica, Molecular y Nuclear and GISC, Universidad Complutense de Madrid - Universidad Complutense de Madrid, 28040 Madrid, Spain

PACS 05.70.-a – Thermodynamics
PACS 05.20.-y – Classical statistical mechanics
PACS 89.70.-a – Information and communication theory

Abstract. - The sum of the average work dissipated plus the information gained during a thermodynamic process with discrete feedback must exceed zero. We demonstrate that the minimum value of zero is attained only by feedback-reversible processes that are indistinguishable from their time-reversal, thereby extending the notion of thermodynamic reversibility to feedback processes. In addition, we prove that in every realization of a feedback-reversible process the sum of the work dissipated and change in uncertainty is zero.

Investigations into the thermodynamic implications of feedback have a long history [1,2], especially with regard to the relationship between information acquisition and thermodynamic quantities – such as work, heat, and entropy. Still, there is no complete theoretical framework detailing the relationship between feedback and thermodynamics. Interest in developing such a framework – a *thermodynamics of feedback* – has grown recently in part due to experiments on feedback cooling [3,4] and feedback control of nanoparticles [5,6]; experimental, computational, and theoretical studies of feedback driven Brownian ratchets [7–12]; and new theoretical predictions relating dissipation to information [13–23].

Recently, Sagawa and Ueda derived a generalization of the second law of thermodynamics for quantum and classical systems manipulated by one feedback loop [18,19], which subsequently has been verified experimentally [24] and extended to classical systems driven by repeated discrete feedback – implemented through a series of feedback loops initiated at predetermined times – independently by Horowitz and Vaikuntanathan [20], and Fujitani and Suzuki [21]. *The second law of thermodynamics for discrete feedback* states that the average work dissipated $\langle W_d \rangle$ in driving a system with a discrete feedback protocol from one equilibrium state at inverse temperature β to another at the same temperature is related to the microscopic information gained through measurements $\langle I \rangle$ by

$$\beta \langle W_d \rangle + \langle I \rangle \geq 0. \quad (1)$$

Here, $\langle I \rangle$ is the mutual information between the state

of the system and the measurement outcome [18,19,25], and the average work dissipated is the average work $\langle W \rangle$ done in excess of the average free-energy difference $\langle \Delta F \rangle$: $\langle W_d \rangle = \langle W \rangle - \langle \Delta F \rangle$. The nomenclature derives from the observation that in the absence of feedback [$\langle I \rangle = 0$], eq. (1) reduces to a statement of the second law of thermodynamics: $\langle W_d \rangle \geq 0$.

The second law of the thermodynamics plays a central role within the framework of thermodynamics. Besides restricting the realizability of thermodynamic processes, it distinguishes particular processes that produce no entropy. In macroscopic systems, such processes are called reversible, because the sequence of equilibrium states visited by the system during the process can be traversed both forwards and backwards [26]. At the microscopic level the connection between entropy production and reversibility must be interpreted statistically and is quantified mathematically by the distinguishability of a thermodynamic process from its time-reversal [22,23,27]. Like the second law of thermodynamics, eq. (1) singles out particular processes that have no dissipation [$\beta \langle W_d \rangle + \langle I \rangle = 0$]. Because such processes are important in thermodynamics, it is of value to characterize them in the presence of feedback; especially since they are optimal processes that most efficiently convert information into work. Thus, in this letter we analyze those processes that saturate the bound in eq. (1) [$\beta \langle W_d \rangle + \langle I \rangle = 0$], and we demonstrate that such processes satisfy a new, distinct criterion, similar to thermodynamic reversibility, which we call *feedback reversibility*. Intuition from the second law of thermody-

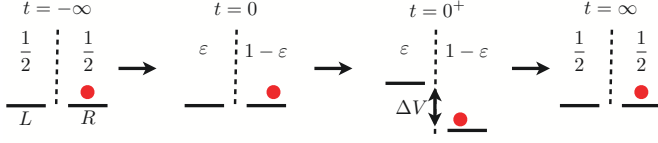


Fig. 1: Illustration of the feedback protocol driving a particle (dot) that makes thermally activated transitions between two states L and R . From left to right is the evolution of the relative heights of the energy levels of the two states (horizontal bars) at four times, $t = -\infty, 0, 0^+$ and ∞ , during the protocol $\Lambda^R = \{\lambda_t^R\}_{t=0}^\infty$ associated to measuring the particle to be in state R at $t = 0$. Above each state is the probability to be in that state conditioned on implementing protocol Λ^R , $\rho_t(k|\Lambda^R)$, $k = L, R$.

namics might lead one to naively expect that the bound in eq. (1) can always be reached as long as the process is sufficiently slow; however, this is generally not true, as we will demonstrate. To saturate eq. (1) one must balance $\langle W_d \rangle$ and $\langle I \rangle$, which requires adjusting the driving protocol, the type of measurement, and the measurement error.

Before considering the most general situation, let us first analyze a simple toy model of feedback, inspired by the Szilard engine [1, 2]. We will demonstrate that by adjusting the measurement error we can saturate eq. (1) even when the process is not adiabatically slow. Consider a particle weakly coupled to a thermal bath at inverse temperature $\beta = 1$ that makes thermally activated jumps between two states, labeled left (L) and right (R). The state energies $E_i(\lambda)$, $i = L, R$, depend on a vector of external parameters λ which we vary using feedback during a time interval $t \in (-\infty, \infty)$. Initially at $t = -\infty$ the particle is in equilibrium with the energies equal. From $t = -\infty$ to 0 , the system evolves freely. At $t = 0$, we measure the particle's state, misidentifying it with error ε , *i.e.*, the probability to measure the particle in R (L) given it is in L (R) is $P(R|L) = \varepsilon$ [$P(L|R) = \varepsilon$]. Immediately after the measurement at $t = 0^+$, we vary the external parameters according to a protocol that depends on the measurement outcome: if the particle is measured to be in state R (L), we initiate protocol $\Lambda^R = \{\lambda_t^R\}_{t=0}^\infty$ (Λ^L), depicted in fig. 1, by instantaneously lowering the right (left) energy level by $\Delta V/2$, while simultaneously raising the left (right) energy level by $\Delta V/2$; so that the energy difference between the two states is ΔV . Finally, from $t = 0^+$ to ∞ , we quasi-statically return the energy levels to their original values. (Similar protocols were utilized to study dissipation in nonselective quantum measurement [28] and the achievability of the equality in eq. (1) for quantum feedback given a particular measurement protocol [29].)

Figure 1 contains a depiction of a realization of this process where the protocol Λ^R is executed. At the four times $t = -\infty, 0, 0^+$, and ∞ along the process, we depict the relative heights of the energy levels of the two states. Above each energy level is the probability at time t to be in that

state *conditioned* on implementing protocol Λ^R , $\rho_t(k|\Lambda^R)$, $k = L, R$. Initially the particle is in equilibrium with energy levels equal, therefore $\rho_{-\infty}(R|\Lambda^R) = \rho_{-\infty}(L|\Lambda^R) = 1/2$. At $t = 0$, we measure the particle to be in state R ; consequently, $\rho_0(R|\Lambda^R) = 1 - \varepsilon$ and $\rho_0(L|\Lambda^R) = \varepsilon$ reflecting that with probability ε protocol Λ^R is mistakenly implemented when the particle is in state L . Immediately after the measurement we instantaneously change the energy levels. Since this step is infinitely fast, the conditional probability distributions do not vary. Finally, starting at $t = 0^+$ we infinitely slowly return the energy levels to their original configuration, so that at $t = \infty$ the system has returned to its initial equilibrium.

To explore how eq. (1) depends on the error ε , we determine the values of $\langle W_d \rangle$ and $\langle I \rangle$ as functions of ε . First, observe that since each protocol is cyclic the average free-energy difference is zero $\langle \Delta F \rangle = 0$, and the dissipated work equals the work, $\langle W_d \rangle = \langle W \rangle$. Furthermore, the symmetry of Λ^R and Λ^L implies that the average work during each protocol is the same. Thus, we focus on the average work during Λ^R , which we calculate in two steps. First, we compute the average work conditioned on implementing protocol Λ^R during the instantaneous switching of the energy levels at $t = 0^+$:

$$\langle w_1 \rangle_R = \varepsilon \frac{\Delta V}{2} - (1 - \varepsilon) \frac{\Delta V}{2}. \quad (2)$$

Second, from time $t = 0^+$ to ∞ the process is quasi-static. Therefore, the average work given Λ^R during this period is the free-energy difference between the equilibrium states when the energy levels differ by ΔV and when the energy levels are equal:

$$\langle w_2 \rangle_R = -\ln \frac{2}{e^{-\Delta V/2} + e^{\Delta V/2}}. \quad (3)$$

Noting that each protocol, Λ^R and Λ^L , occurs with equal probability and that the work during each protocol is the same, we conclude that the total average work equals the sum of eqs. (2) and (3), which, after some algebraic manipulation, can be expressed as

$$\langle W \rangle = -\ln 2 - (1 - \varepsilon) \ln \frac{e^{\Delta V}}{1 + e^{\Delta V}} - \varepsilon \ln \frac{1}{1 + e^{\Delta V}}. \quad (4)$$

The mutual information $\langle I \rangle$ [see eq. (11) below] quantifies the reduction in our uncertainty about the microscopic state of the system upon making a measurement. It is defined as the relative entropy between the joint probability distribution of the state of the system $k = L, R$ at the time of measurement ($t = 0$) and the measurement outcome $m = L, R$,

$$\rho_0(k, m) = \begin{cases} \frac{1}{2}(1 - \varepsilon) & k = m \\ \frac{1}{2}\varepsilon & k \neq m \end{cases}, \quad (5)$$

with the product of their respective marginal distributions, $\rho_0(k) = 1/2$ and $P(m) = 1/2$ [25]. Therefore, the

mutual information reads:

$$\begin{aligned}
 \langle I \rangle &= \sum_{k,m} \rho_0(k,m) \ln \frac{\rho_0(k,m)}{\rho_0(k)P(m)} \\
 &= (1-\varepsilon) \ln \left[\frac{(1-\varepsilon)/2}{1/4} \right] + \varepsilon \ln \left[\frac{\varepsilon/2}{1/4} \right] \\
 &= \ln 2 + (1-\varepsilon) \ln(1-\varepsilon) + \varepsilon \ln \varepsilon.
 \end{aligned} \quad (6)$$

In fig. 2 we plot $\langle W \rangle$ [eq. (4)], $\langle I \rangle$ [eq. (6)], and their sum

$$\langle W \rangle + \langle I \rangle = (1-\varepsilon) \ln \frac{(1-\varepsilon)(1+e^{\Delta V})}{e^{\Delta V}} + \varepsilon \ln [\varepsilon(1+e^{\Delta V})] \quad (7)$$

as functions of the error ε . For error-free measurements

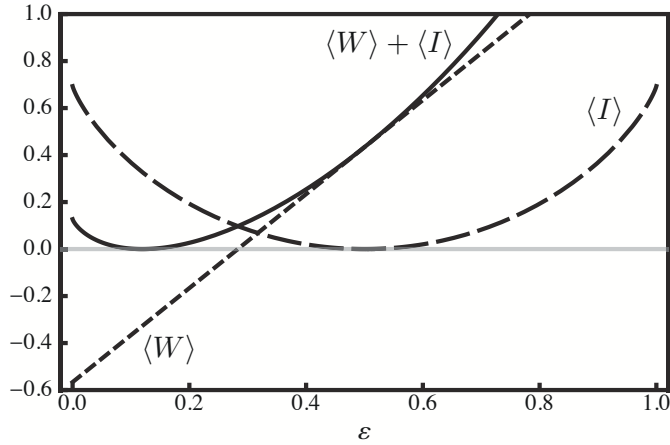


Fig. 2: Plot of $\langle W \rangle$ (dashed), $\langle I \rangle$ (long dashed), and their sum $\langle W \rangle + \langle I \rangle$ (solid) as a function of the measurement error ε for $\Delta V = 2$.

($\varepsilon = 0$), we are able to extract the maximum amount of work [$-\langle W \rangle$ is largest] and obtain the maximum amount of information $\langle I \rangle = \ln 2$. However, $\langle W \rangle + \langle I \rangle > 0$; some of the information is not used to extract work. At the expense of increasing ε , decreasing the amount of work extracted, and decreasing the amount of information gained, we can reach the bound in eq. (1). From eq. (7), we see that the sum $\langle W \rangle + \langle I \rangle$ is zero when ε equals

$$\varepsilon_0 = \frac{1}{1+e^{\Delta V}}. \quad (8)$$

Thus, by adjusting the measurement error with fixed external parameter protocols we can construct a feedback protocol that is not adiabatically slow and satisfies $\beta \langle W_d \rangle + \langle I \rangle = 0$.

A physical interpretation can be given to ε_0 in eq. (8) by considering the time-reversal of the preceding feedback process. The time-reversal of a feedback process is not trivial, since there is no such thing as the time-reversal of a measurement. However, we follow Ref. [20] and construct a distinct thermodynamic process termed the *reverse* process, which will act as the time-reversed feedback process. The reverse process begins by randomly

selecting a protocol, Λ^L or Λ^R , according to the probability that the respective measurement outcome, L or R , occurs during feedback: $P(L) = P(R) = 1/2$. The particle is then driven away from equilibrium by using the time-reversal of the selected protocol, $\tilde{\Lambda}^k = \{\lambda_{-t}^k\}_{t=-\infty}^0$, $k = L, R$. Now, imagine randomly selecting protocol Λ^R , and consider the evolution of the system conditioned on executing the time-reversed protocol $\tilde{\Lambda}^R$, depicted in fig. 3. Initially the particle is in equilibrium with ener-

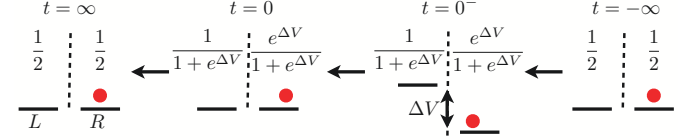


Fig. 3: Illustration of the reverse process at four times, $t = -\infty, 0^-, 0$ and ∞ (right to left), along the protocol $\tilde{\Lambda}^R$. Above each state is the conditional probability distribution $\tilde{\rho}_t(k|\tilde{\Lambda}^R)$, $k = L, R$.

gies equal. As a result, the probabilities to find the particle in states L and R at $t = -\infty$ in the reverse process *conditioned* on the protocol $\tilde{\Lambda}^R$ are initially equal: $\tilde{\rho}_{-\infty}(R|\tilde{\Lambda}^R) = \tilde{\rho}_{-\infty}(L|\tilde{\Lambda}^R) = 1/2$. From $t = -\infty$ to 0^- , the right energy level is quasi-statically lowered by $\Delta V/2$ while the left energy level is raised by $\Delta V/2$. Thus, at $t = 0^-$ the conditional probability distribution is a Boltzmann distribution: $\tilde{\rho}_{0^-}(L|\tilde{\Lambda}^R) = 1/(1+e^{\Delta V})$ and $\tilde{\rho}_{0^-}(R|\tilde{\Lambda}^R) = e^{\Delta V}/(1+e^{\Delta V})$. At $t = 0$, the energy levels are instantaneously returned to their original equal values; the conditional probability distribution does not change. Finally, from $t = 0$ to ∞ the energy levels are held fixed, while the system relaxes back to its initial equilibrium.

Comparing figs. 1 and 3, we see that when $\varepsilon = \varepsilon_0$ [eq. (8)] the conditional probability distributions in the feedback and the reverse processes are equal at the main stages of the process:

$$\rho_t(k|\Lambda^R) = \tilde{\rho}_{-t}(k|\tilde{\Lambda}^R), \quad (9)$$

for $k = L, R$. The same holds for Λ^L . Below we demonstrate that this equality holds at any time t [see eq. (18)]. Moreover, since each protocol is implemented with equal likelihood in both the original feedback process and the reverse process, the joint distributions of states and protocols for the feedback and reverse processes are equal for an optimal feedback process: $\rho_t(k, \Lambda^m) = \tilde{\rho}_{-t}(k, \tilde{\Lambda}^m)$, $k, m = L, R$. Notice that the reverse and feedback processes are of a different nature: there are no measurements in the reverse process. Despite this difference, identifying the reverse processes as the time-reversed feedback processes is consistent with the thermodynamic principle linking reversibility and dissipation: the process with least dissipation – the protocol which saturates eq. (1) – corresponds to the situation where the feedback process is indistinguishable from its time-reversal (from the reverse

process). Consequently, we call these processes *feedback reversible* or simply *reversible*.

We now demonstrate that this conclusion holds quite generally as a consequence of the detailed fluctuation theorem for discrete feedback [eq. (13) below] [20]. This theorem relates the fluctuations in two thermodynamic processes related by time-reversal: the forward and the reverse processes. Our analysis begins by sketching a derivation of eq. (1) based on the detailed fluctuation theorem for discrete feedback originally presented in Ref. [20]. For clarity of exposition, we limit our discussion to feedback processes with only one feedback loop initiated at the beginning of the process. All our conclusions can be generalized to situations with repeated discrete feedback.

Consider a classical system, whose position in phase space (or microscopic configuration) at time t is z_t . This system is driven away from equilibrium at inverse temperature β from $t = 0$ to τ using one feedback loop as follows: at $t = 0$ a physical observable M is measured with outcomes m that occur with probability $P(m|z_0)$, conditioned on the system's state at the time of measurement z_0 . Based on the outcome of this measurement, the system is driven by varying a set of external parameters λ with time using the protocol $\Lambda^m = \{\lambda_t^m\}_{t=0}^\tau$ from $\lambda_0^m = A$ to $\lambda_\tau^m = B^m$. Finally, at time $t = \tau$ the external parameters are held fixed at $\lambda_\tau^m = B^m$ while the system relaxes back to equilibrium at inverse temperature β . Repeatedly executing this sequence of actions – each time equilibrating the system, driving the system using feedback, and then re-equilibrating the system – generates an ensemble of realizations of the *forward* process. In each realization, thermal fluctuations will cause the system to trace out a different microscopic trajectory through phase space $\gamma = \{z_t\}_{t=0}^\tau$. The joint probability to observe γ with Λ^m is $\mathcal{P}[\gamma; \Lambda^m]$. The work dissipated during this realization,

$$W_d[\gamma; \Lambda^m] = W[\gamma; \Lambda^m] - \Delta F[\Lambda^m], \quad (10)$$

is the work $W[\gamma; \Lambda^m]$ done in excess of the free-energy difference $\Delta F[\Lambda^m]$. Here, $\Delta F[\Lambda^m]$ depends on the executed protocol, because the final external parameter value is a function of the measurement, $\lambda_\tau^m = B^m$. Moreover, measurements made upon initiating the feedback loop change our uncertainty about the microscopic state of the system by an amount [20]

$$I[\gamma; \Lambda^m] = \ln \left[\frac{P(m|z_0)}{P(m)} \right], \quad (11)$$

where $P(m)$ is the (unconditional) probability to observe outcome m when measuring the physical observable M . The average of I over many realizations is the mutual information that appears in eq. (1).

The reverse process is defined as in our previous example. In the reverse process no measurements are made. Instead, we drive the system away from equilibrium using a *reverse* protocol $\tilde{\Lambda}^m = \{\tilde{\lambda}_t^m\}_{t=0}^\tau$ with $\tilde{\lambda}_t^m = \lambda_{\tau-t}^m$, which is selected randomly with probability $\tilde{\pi}[\tilde{\Lambda}^m]$ defined to be

equal to the probability to implement the forward protocol $\lambda_t^m = \tilde{\lambda}_{\tau-t}^m$ in the forward process $\pi[\Lambda^m]$:

$$\tilde{\pi}[\tilde{\Lambda}^m] = \pi[\Lambda^m] = \int d\gamma \mathcal{P}[\gamma; \Lambda^m], \quad (12)$$

where $d\gamma$ is a measure on microscopic trajectory space. After selecting $\tilde{\Lambda}^m$, the system is equilibrated with a thermal bath at inverse temperature β with external parameters fixed at $\tilde{\lambda}_0^m = \lambda_\tau^m = B^m$. Observe that the initial equilibrium distribution of the reverse process corresponds to external parameter value B^m and depends on which protocol $\tilde{\Lambda}^m$ is implemented [20]. From $t = 0$ to τ the external parameters are varied according to the reverse protocol $\tilde{\Lambda}^m = \{\tilde{\lambda}_t^m\}_{t=0}^\tau$. At $t = \tau$, the external parameters are fixed to $\tilde{\lambda}_\tau^m = \lambda_0 = A$ while the system relaxes back to equilibrium. For every microscopic trajectory of the forward process $\gamma = \{z_t\}_{t=0}^\tau$, there is a *conjugate* reverse trajectory $\tilde{\gamma} = \{\tilde{z}_t\}_{t=0}^\tau$, where $\tilde{z}_t = z_{\tau-t}^*$ and z^* denotes momentum reversal. The probability to observe reverse trajectory $\tilde{\gamma}$ and reverse protocol $\tilde{\Lambda}^m$ in the reverse process is $\tilde{\mathcal{P}}[\tilde{\gamma}; \tilde{\Lambda}^m]$.

Having introduced notation and presented the definitions of the forward and reverse processes, we now state the detailed fluctuation theorem [20]:

$$\frac{\mathcal{P}[\gamma; \Lambda^m]}{\tilde{\mathcal{P}}[\tilde{\gamma}; \tilde{\Lambda}^m]} = e^{\beta W_d[\gamma; \Lambda^m] + I[\gamma; \Lambda^m]}. \quad (13)$$

Because no measurements are made in the reverse process, there are microscopic trajectories and reverse protocols that occur together in the reverse process whose conjugate microscopic trajectories and conjugate protocols do *not* occur together in the forward process; it is possible for $\tilde{\mathcal{P}} \neq 0$ when $\mathcal{P} = 0$ [20].

Equation (13) implies that the relative entropy, $D(f||g) = \int dx f(x) \ln(f(x)/g(x))$, between \mathcal{P} and $\tilde{\mathcal{P}}$ is [20]

$$D(\mathcal{P}||\tilde{\mathcal{P}}) = \beta \langle W_d \rangle + \langle I \rangle, \quad (14)$$

where $\langle \cdot \rangle$ is an ensemble average over realizations of the forward process. $D(\mathcal{P}||\tilde{\mathcal{P}})$ measures the distinguishability of the forward and reverse processes; it is a microscopic measure of the intensity of the “arrow of time” [22, 23].

Equation (1) now follows by exploiting the nonnegativity of the relative entropy ($D \geq 0$) [25] in eq. (14). Moreover, eq. (14) implies that thermodynamic processes for which $\beta \langle W_d \rangle + \langle I \rangle = 0$ are those with $D = 0$, which is true if and only if

$$\mathcal{P}[\gamma; \Lambda^m] = \tilde{\mathcal{P}}[\tilde{\gamma}; \tilde{\Lambda}^m] \quad (15)$$

for all γ and Λ^m [25]. The condition $D = 0$ additionally requires that the supports of \mathcal{P} and $\tilde{\mathcal{P}}$ – the sets of microscopic trajectories and protocols for which \mathcal{P} and $\tilde{\mathcal{P}}$ are nonzero – must be identical: every microscopic trajectory and reverse protocol that occur together in the reverse process have conjugate pairs that occur together in the

forward process. Equation (15) is a microscopic statement of reversibility: the process looks the same forwards and backwards in time, since every realization occurs with equal likelihood in the forward and reverse processes. We conclude that the inequality in eq. (1) is saturated only when eq. (15) is satisfied, that is *only for reversible processes*.

One could in principle consider a “super-system” composed of our system of interest and a feedback mechanism – formed from a controller that manipulates the parameters λ and a memory that records the measurement outcomes. While the explicit implementation of such a set-up with a time-dependent Hamiltonian is a difficult task, it is still instructive to assume that a super-system exists and to compare the feedback reversibility we have introduced with standard thermodynamic reversibility. In this set-up, we manipulate the super-system by varying a collection of parameters external to both the system of interest and the feedback mechanism according to a predetermined cyclic protocol. The work performed by the external parameters along this cycle is responsible for recording the measurement of the system, the control in response to the measurement, and the erasure of the measurement. As shown by Sagawa and Ueda [16], under rather general hypothesis, the work to measure $\langle W_{\text{meas}} \rangle$ plus the work to erase $\langle W_{\text{eras}} \rangle$ is bound by $\beta(\langle W_{\text{eras}} \rangle + \langle W_{\text{meas}} \rangle) \geq \langle I \rangle$. Combining this bound with equation (1), we find that the total work performed on the super-system is

$$\langle W_d \rangle + \langle W_{\text{eras}} \rangle + \langle W_{\text{meas}} \rangle \geq 0, \quad (16)$$

which is the second law of thermodynamics applied to the super-system when the feedback mechanism has gone through a cycle. To achieve the equality in equation (16), and consequently in (1), one would need thermodynamic reversibility in the full super-system – that is the thermodynamic process must be indistinguishable from its time-reversal in the phase space of the super-system. However, such a process is not quasi-static, because during the writing and erasure of the memory the super-system is not in equilibrium. We have shown that to achieve the optimal feedback control, equality in (1), one only needs feedback reversibility in the system, while the controller can be out of equilibrium.

From eq. (15) follows another useful microscopic statement of reversibility in terms of the phase space densities conditioned on the executed protocol [eq. (18) below]. Recall that in our construction of the reverse process, the probability to execute Λ^m in the forward process $\pi[\Lambda^m]$ equals the probability to use the conjugate reverse protocol $\tilde{\Lambda}^m$ in the reverse process $\tilde{\pi}[\tilde{\Lambda}^m]$ [eq. (12)]. Thus dividing both sides of eq. (15) by $\pi[\Lambda^m] = \tilde{\pi}[\tilde{\Lambda}^m]$ we find that, for a reversible process, the statistics of the trajectories *conditioned* on the protocol executed in the forward and reverse processes are also indistinguishable:

$$\mathcal{P}[\gamma|\Lambda^m] = \tilde{\mathcal{P}}[\tilde{\gamma}|\tilde{\Lambda}^m]. \quad (17)$$

Furthermore, by integrating these conditional trajectory distributions over all trajectories that pass through $z_t = \tilde{z}_{\tau-t}^*$, we find that the conditional phase space densities must also be equal:

$$\rho(z_t|\Lambda^m) = \tilde{\rho}(\tilde{z}_{\tau-t}|\tilde{\Lambda}^m), \quad (18)$$

which is the general expression corresponding to eq. (9).

Equation (18) is valid at all times during a feedback-reversible process. In particular, it implies that in our preceding example eq. (9) is true even during the time interval $t \in (-\infty, 0)$ prior to the measurement. The reversibility of the process during this interval may be surprising at first, since during this period the system is freely evolving with the external parameters held fixed. To gain further insight, consider a special case of our example with no measurement error, $\varepsilon = 0$. In the forward process during the interval $t \in (-\infty, 0)$, $\rho_t(k|\Lambda^R)$ represents the evolution of the (probabilistic) state of the system conditioned on the *future* measurement outcome being R at $t = 0$. Whereas during this interval, $\tilde{\rho}_{-t}(k|\tilde{\Lambda}^R)$ is a relaxation from a non-equilibrium distribution $\tilde{\rho}_0(k|\tilde{\Lambda}^R) = \delta_{k,R}$ to equilibrium $\tilde{\rho}_\infty(k|\tilde{\Lambda}^R) = 1/2$. The reversibility condition eq. (18) [eq. (9)] states that these two processes are identical upon time reversal. In fig. 4, we illustrate this reversibility by depicting the evolution of an ensemble of systems during $t \in (-\infty, 0)$. The ensemble is in equilibrium from $t = -\infty$ to 0, but the sub-ensemble of systems that are in the right (left) state at the time of measurement $t = 0$ exhibits an evolution that is identical to the time-reversal of the corresponding relaxation. The underlying reason is microscopic reversibility: the probability in equilibrium to observe a microscopic trajectory and its time reversal are equal [30]. As a consequence, the probability to observe a microscopic trajectory conditioned on being at a particular state in the future, say z' , is the same as the probability to observe the time reversal of this trajectory while the system relaxes to equilibrium when initially at z' .

Finally, we note that in *every* realization of a reversible process $\beta W_d + I = 0$, which follows by substituting the reversibility condition eq. (15) into the nonequilibrium detailed fluctuation theorem [eq. (13)]. Although $\beta W_d + I = 0$ for every realization of a reversible process, the value of W_d (I) may differ in each realization.

We have demonstrated that those feedback processes that most efficiently use information to extract work [that saturate eq. (1)] are feedback reversible – they are indistinguishable from their time-reversal, thereby extending the concept of thermodynamic reversibility to feedback processes. Like reversible thermodynamic processes, feedback-reversible processes are ideal for experimentally measuring free-energy differences. When all external parameter protocols end at the same value $\lambda_\tau^m = B$ – independent of m – the sum of the work and information along a feedback-reversible process equals the free-energy difference. Feedback-reversible processes can also be used

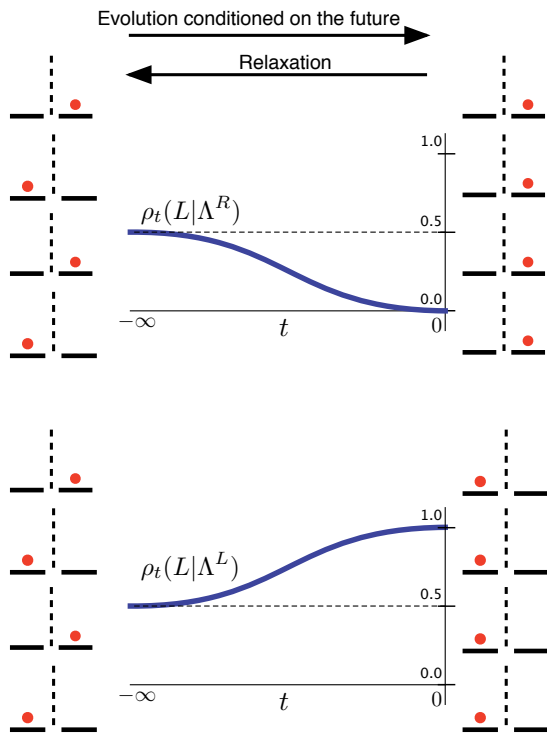


Fig. 4: During the time interval $t = -\infty$ to 0 , the external parameters are constant with time; therefore, an ensemble of systems initially in equilibrium remains in equilibrium. However, the sub-ensemble of systems that at $t = 0$ are measured to be in the right (left) state – depicted here as the four upper (lower) copies of the system – does vary in time. The probability in this sub-ensemble to be in state L at time t , $\rho_t(L|\Lambda^R)$ [$\rho_t(L|\Lambda^L)$], is plotted as a function of time during the interval $t \in (-\infty, 0)$, assuming error-free measurement ($\varepsilon = 0$). In a reversible process, the time reversal of this evolution is identical to a relaxation to equilibrium from a nonequilibrium state given by $\rho_0(L|\Lambda^R) = 0$ [$\rho_0(L|\Lambda^L) = 1$].

to estimate the information gain by measuring the work dissipated. In our illustrative example, if we were to measure the work as a function of ΔV its minimum value would equal $\langle I \rangle$ [eq. (4)]. In general, a measurement of the work dissipated gives a lower bound on the information according to eq. (1). Furthermore, our work gives an indication on how to design feedback-reversible protocols for a given measurement scheme, which may be a difficult task [31, 32]. The protocols must be such that eq. (18) holds at every instant. In particular, immediately after a measurement with outcome m , the conditioned phase space density must be identical to the conditioned density prepared by the protocol λ_t^m run in reverse. This identity is the key ingredient for designing an optimal protocol.

The authors acknowledge financial support from Grant MOSAICO (Spanish Government) and MODELICO (Comunidad de Madrid).

REFERENCES

- [1] MARUYAMA K., NORI F. and VEDRAL V., *Rev. Mod. Phys.*, **81** (2009) 1.
- [2] LEFF H. S. and REX A. F., (Editors) *Maxwell's Demon: Entropy, Information, Computing* (Princeton University Press, New Jersey) 1990.
- [3] COHADON P. F., HEIDMANN A. and PINARD M., *Phys. Rev. Lett.*, **83** (1999) 3174.
- [4] LIANG S., MEDICH D., CZAJKOWSKY D. M., SHENG S., YUAN J.-Y. and SHAO Z., *Ultramicroscopy*, **84** (2000) 118.
- [5] COHEN E. G. D. and MAUZERALL D., *Mol. Phys.*, **103** (2005) 2923.
- [6] COHEN A. and MOERNER W., *Proc. Natl. Acad. Sci.*, **103** (2006) 4362.
- [7] CAO F. J., DINIS L. and PARRONDO J. M. R., *Phys. Rev. Lett.*, **93** (2004) 040603.
- [8] DINIS L., PARRONDO J. M. R. and CAO F. J., *Europhys. Lett.*, **71** (2005) 536.
- [9] FEITO M. and CAO F. J., *Eur. Phys. J. B*, **59** (2007) 63.
- [10] LOPEZ B. J., KUWADA N. J., CRAIG E. M., LONG B. R. and LINKE H., *Phys. Rev. Lett.*, **101** (2008) 220601.
- [11] CRAIG E. M., LONG B. R., PARRONDO J. M. R. and LINKE H., *Europhys. Lett.*, **81** (2008) 10002.
- [12] CAO F. J., FEITO M. and TOUCHETTE H., *Physica A*, **338** (2009) 112.
- [13] KIM K. H. and QIAN H., *Phys. Rev. Lett.*, **93** (2004) 120602.
- [14] KIM K. H. and QIAN H., *Phys. Rev. E*, **75** (2007) 022102.
- [15] ALLAHVERDYAN A. E. and SAAKIAN D. B., *Europhys. Lett.*, **81** (2008) 30003.
- [16] SAGAWA T. and UEDA M., *Phys. Rev. Lett.*, **102** (2009) 250602.
- [17] PONMURUGAN M., *Phys. Rev. E*, **82** (2010) 031129.
- [18] SAGAWA T. and UEDA M., *Phys. Rev. Lett.*, **100** (2008) 080403.
- [19] SAGAWA T. and UEDA M., *Phys. Rev. Lett.*, **104** (2010) 090602.
- [20] HOROWITZ J. M. and VAIKUNTANATHAN S., *Phys. Rev. E*, **82** (2010) 061120.
- [21] FUJITANI Y. and SUZUKI H., *J. Phys. Soc. Jap.*, **79** (2010) 104003.
- [22] KAWAI R., PARRONDO J. M. R. and VAN DEN BROECK C., *Phys. Rev. Lett.*, **98** (2007) 080602.
- [23] PARRONDO J. M. R., VAN DEN BROECK C. and KAWAI R., *New J. Phys.*, **11** (2009) 073008.
- [24] TOYABE S., SAGAWA T., UEDA M., MUNIYUKI E. and SANO M., *Nature Phys.*, **6** (2010) 988.
- [25] COVER T. M. and THOMAS J. A., *Elements of Information Theory* 2nd Edition (Wiley-Interscience) 2006.
- [26] CALLEN H. B., *Thermodynamics and an introduction to thermostatistics, 2nd edition* (John Wiley and Sons, New York) 1985.
- [27] HOROWITZ J. M. and JARZYNSKI C., *Phys. Rev. E*, **79** (2009) 021106.
- [28] EREZ N., *Thermodynamics of quantum measurement* arxiv:1011.1020v1 (2010).
- [29] JACOBS K., *Phys. Rev. A*, **80** (2009) 012322.
- [30] CROOKS G. E., *Phys. Rev. E*, **61** (2000) 2361.
- [31] KIM S. W., SAGAWA T., DE LIBERATO S. and UEDA M., *Phys. Rev. Lett.*, **106** (2011) 070401.

- [32] ABREU D. and SEIFERT U., *Europhys. Lett.* , **94** (2011) 10001.